BIOMETRICS , 1–6

FDR Control with Pseudo-Gatekeeping Based on a Possibly Data Driven Order of the Hypotheses

A. Farcomeni*

Department of Public Health and Infectious Diseases
Sapienza - University of Rome, Piazzale Aldo Moro, 5, 00185 Rome, Italy
**email:* alessio.farcomeni@uniroma1.it

and

L. Finos *
Department of Statistical Sciences
University of Padua, Via Cesare Battisti 241, 35121 Padova, Italy
**email:* livio@stat.unipd.it

SUMMARY: We propose a multiple testing procedure controlling the false discovery rate. The procedure is based on a possibly data driven ordering of the hypotheses, which are tested at the uncorrected level q until a suitable number is not rejected. When the order is data driven, larger effect sizes are considered first, therefore selecting more interesting hypotheses with larger probability. The proposed procedure is valid under independence for the test statistics. We also propose a modification which makes our procedure valid under arbitrary dependence. It is shown in simulation that we compare particularly well when the sample size is small. We conclude with an application to identification of molecular signatures of intracranial ependymoma. The methods are implemented in an R package (`someMTP`), freely available on CRAN. This paper includes supplementary material online.

KEY WORDS: Data driven order; False Discovery Rate; Multiple testing.

1. Introduction

Many modern applications of statistical inference, especially in genomics and neuroscience involve simultaneously testing many hypotheses. Controlling the False Discovery Rate (FDR) in these applications is an appealing approach since it usually provides a small proportion of false rejections among the selected hypotheses, providing a good balance between Type I and Type II errors. The FDR can be loosely defined as the expected proportion of false discoveries, if any. Benjamini and Hochberg (1995) propose to control the FDR at a pre-specified level q using a step-up procedure. The same procedure was shown to control the FDR under positive regression dependency on subsets (PRDS) and extended to arbitrary dependence by Benjamini and Yekutieli (2001). Many works, which we do not attempt to review, have been later devoted to the operating characteristics and properties of the latter procedures and some of their variants. For reviews in the area of multiple testing refer to Dudoit et al. (2003) and to Farcomeni (2008).

In this paper we propose an ordered procedure for FDR control with no threshold correction. Hypotheses are tested at level q along a given order, and the procedure is stopped once a certain number of p-values have been found above q. This strategy can be regarded as pseudo-gatekeeping (e.g., Dmitrienko et al. (2008)), as each sequence of hypotheses along the order represents a gate which should be passed (i.e., there should be enough rejections) in order to allow the user to verify the following hypothesis. A careful choice of the number of required rejections allows for control of the FDR at level q. The order of the hypotheses can be pre-specified, or can be data driven. In the latter case, with our proposed data driven ordering hypotheses with larger effect sizes are more likely to appear at the top of the list; thus more likely preferring biologically significant hypotheses over the others even if they may have a slightly larger p-value. Furthermore, the proposed procedures can be seen in simulation to compare particularly well in terms of power. In this paper we measure power as the proportion of correctly rejected hypotheses over the number of false nulls. We will see that our procedure often yields 2 to 5 times the proportion of correctly rejected hypotheses with respect to competitors when the sample size is small.

In our motivating application, a microarray-based comparative genomic hybridization (aCGH) was used to compare infratentorial versus supratentorial tumors. The same data were analyzed in Modena et al. (2006) from a different perspective. The sample size of $n = 22$ is relatively low given that the signal is very weak and sparse among probes (i.e. variables). Commonly used procedures have low power in these cases, and in fact with these data they lead to no rejections for significance levels of 1% and 5%.

The idea of testing along a given order is not new. Maurer et al. (1995) propose to test hypotheses along a pre-specified order, and to stop when the first p-value is found above q, showing that this procedure controls the Familywise error rate (FWER). Kropf and Läuter (2002); Kropf et al. (2004) use the latter procedure with a data driven order for controlling the FWER. The idea of allowing for more than one p-value above the significance level is used in Hommel and Kropf (2005) to enhance power when controlling the FWER, and in Finos and Farcomeni (2011) for controlling the k-FWER, i.e., the probability of k or more false rejections. We propose here an FDR controlling procedure in the spirit of the previous papers. When hypotheses have a pre-specified ordering, p-values can arise from any kind of hypothesis testing. When the order of the hypotheses is data driven, we have the additional assumption that p-values are based on a certain general linear model (i.e., we restrict to continuous outcomes). The p-values can then arise, under a data driven ordering, from one or two sample t-testing, analysis of variance, analysis of covariance, regression. One can adjust the p-values for confounders and non-parametric approaches can be accommodated via the rank transformation (Conover and Iman, 1982). Similar results on data driven ordering are shown for instance in Läuter et al. (1998), later extended to the more general case of covariates in Finos (2011). Here we propose a much simpler proof based on the Basu theorem, and further generalize previous results to a wider class of test statistics.

We show our method is valid under independence. We then propose a modification for arbitrary dependence which will be seen in simulation to have approximately the same power as the Benjamini and Hochberg (1995) procedure, which is valid only under PRDS assumptions.

Our methods are implemented in an `R` package (`someMTP`), which can be freely downloaded from `CRAN` at `http://cran.r-project.org/`.

The rest of the paper is as follows: in the next section we outline the setup, discuss situations of a priori ordered hypotheses, and show how to control the FDR in the latter case. In Section 3 we restrict to tests based on linear models and show that in that case one can use a data driven ordering which enhances power without leading to violation of the nominal error rate. In Section 4 we illustrate and compare our approach through simulations and in Section 5 on the original data example.

2. Controlling the False Discovery Rate with ordered hypotheses

Suppose we are testing m hypotheses simultaneously, let $H_0(i)$ be the null hypothesis associated with the i-th test, $H_1(i)$ the corresponding alternative, and p_i the corresponding p-value. In this paper we only work with simple null hypotheses. We assume p-values are *valid*, i.e., upper bounded by the uniform distribution: $\Pr(p_i \leqslant t | H_0(i)) \leqslant t$. This accommodates also testing with discrete data. Table 1 summarizes the possible outcomes of multiple testing: M_0 of the m nulls are true, M_1 are false, and R are rejected. With $N_{1|0}$ and $N_{1|1}$ we denote the number of null hypotheses rejected among the true and false nulls, respectively. Similarly, $N_{0|0}$ and $N_{0|1}$ denote the number of null hypotheses retained among the true and false nulls, respectively. Note that $N_{1|0}$ and $N_{0|1}$ give the actual number of Type I (false positive) and II (false negative) errors, respectively.

[Table 1 about here.]

The FDR is defined as $FDR = E\left[\frac{N_{1|0}}{R+I(R=0)}\right]$, where $I(C)$ is the indicator function for condition C. The FDR is then based on the proportion of type I errors *among the number of rejected hypotheses*, and we want to guarantee that $FDR < q$, for a certain $q \in (0, 1)$.

Throughout we adopt the notation $p_{(i)}$ to denote the i-th ordered p-value, with $p_{(0)} = 0$ and $p_{(m+1)} = 1$, where the ordering is *not* determined in this paper by the rank statistics of p-values, but is pre-specified or data driven according to criteria we specify below. More formally, we adopt the following assumption on p-value ordering:

$$\Pr(\bigcap_j p_j \leqslant t_j) = \Pr(\bigcap_j p_{(j)} \leqslant t_j). \qquad (1)$$

Assumption (1) is a consequence of any *a priori* or data independent ordering of the hypotheses. Note that when hypotheses are ordered according for instance to ranks of p-values as in the Benjamini and Hochberg (1995) procedure, (1) is obviously violated. Situations in which (1) holds arise when hypotheses are generated (and tested) sequentially, in dose-response studies, in toxicity studies, in observational studies when comparing a treatment to more than one type of control (Rosenbaum, 2008), and in other cases. Further, in clinical trials we may have many endpoints ranked before the experiment. Any *a priori* knowledge about the problem can be exploited to drive the ordering of the hypothesis. Hence, the ordering can be derived also by results of previous studies (i.e. independent data) or by other scientific deductions. Among the many references, see for instance Marcus et al. (1976); Hsu and Berger (1999); Maurer et al. (1995); Strassburger et al. (2007). Our procedure allows to test hypotheses sequentially: not all data need to be collected before testing starts, and one could in principle collect new data for testing only if the algorithm has not stopped yet. Unlike many other multiple testing procedures, in our case the number of tests does not need to be known in advance.

2.1 *Independent or Positive Regression Dependent p-values*

In this section we propose our algorithm for ordered FDR control. Hypotheses are verified along the pre-specified order, i.e., we verify whether $p_{(i)} < q$ beginning with $i = 1$. At each step we either stop the procedure (and do not consider the remaining hypotheses) or we continue testing. More precisely, at each step we check that there are less than $J(i, q) = i(1 - q)/(2 - q)$ p-values above q among the first i. Consequently, if $p_{(1)} > q$, the procedure stops with no rejections. Otherwise, p-values are considered one at a time and the procedure stops when the number of p-values found above q is larger than $i(1 - q)/(2 - q)$; or

when all m hypotheses have been reached. A summary is given in Algorithm 1. In Theorem 1 below we formally show that the proposed Algorithm 1 leads to control of the FDR.

Algorithm 1 Sequential procedure for pre-specified ordering of the hypotheses

- Let $i = 1$, $R(1) = I(p_{(1)} < q)$
- while ($R(i) > i - J(i,q)$ and $i \leqslant m$) **do** $i = i + 1$, $R(i) = \sum_{j=1}^{i} I(p_{(j)} < q)$.
- Let $u = i - 1$.
- If $u > 0$, reject all hypotheses for $i = 1, \ldots, u$ corresponding to $p_{(i)} < q$. Do not reject hypotheses for $i > u$ even if $p_{(i)} < q$.

Note that by definition of $J(i,q)$ the procedure is very strict for the first hypotheses (e.g., a single p above q among the first four suffices to stop the procedure in many cases). On the other hand, when a large i has been reached an additional false rejection is not expected to raise the error rate substantially and therefore the i-th hypothesis will be tested if about 50% of the preceding hypotheses are below q. This raises questions about the variability of the error measure, which are tackled in the simulations in the Web Appendix.

We now show that the procedure defined in Algorithm 1 indeed controls the FDR at level q. To this aim let us first prove a key inequality in Lemma 1. Lemma 1 generalizes an inequality obtained in Finos and Farcomeni (2011), Theorem 1, for $J(i,q)$ not depending on i.

LEMMA 1: *Denote with $N_{1|0}(i)$ the number p-values reached by the algorithm stopping at the i-th step, that correspond to true null hypotheses and are below q. Assume also that p-values independent and valid. Then, for any pre-specified sequence $J(i,q)$ that is non-decreasing in i, the probability of k or more type I errors before the $J(i,q)$-th jump is bounded by:*

$$\Pr(N_{1|0}(i) \geqslant k) \leqslant 1 - F_{Bneg(J(i,q),1-q)}(k-1), \quad (2)$$

where $F_{Bneg(J(i,q),1-q)}(\cdot)$ denotes the CDF of a negative binomial with parameters $J(i,q)$ and $1-q$.

We are now ready to show our main result.

THEOREM 1: *Under the assumptions of Lemma 1, the procedure in Algorithm 1 controls the FDR for any prefixed $0 < q < 1$.*

Proofs of Lemma 1 and Theorem 1 are given in Web Appendix.

2.2 *General dependence*

When PRDS of the p-values can not be assumed, a slightly more conservative procedure is given by the following theorem:

THEOREM 2: *Assume that p-values are valid but otherwise arbitrarily dependent. If q is replaced with $q_D = q/\sum_{j=1}^{m}(2-q)/(j+1)$ in Algorithm 1, the same procedure does control FDR at level q.*

A proof of Theorem 2 is given in Web Appendix.

We have shown that FDR control under arbitrary dependence can be achieved by slightly inflating the single inference threshold for the p-values. This of course leads to a slight loss of power with respect to the procedure valid under independence. We will compare power in the simulations below.

3. Data driven order of the hypotheses

The order of the hypotheses can not be chosen *a posteriori* in general. That would in fact be data snooping. Consider for instance the extremal case in which the order of the hypotheses is chosen so that p-values are sorted from the smallest to the largest: all p-values below q would be rejected, and control of the FDR obviously lost. In most cases, on the other hand, there is no natural ordering of the hypotheses. We propose in this section a special data driven ordering which can be proved not to inflate the error rate. Further, our data driven ordering procedure makes it more likely for hypotheses with a large effect size to appear at the beginning of the list, thus rejecting more interesting hypotheses and relegating less interesting (i.e., smaller effect sizes) to the bottom of the list even if the corresponding p-values are small. See e.g. Kirk (2007) and Finos and Farcomeni (2011) on this issue.

When the order of the hypotheses shall be data driven, there are few additional assumptions we must make: we assume hypotheses arise from a linear model on a numerical response, with Gaussian error term. More precisely, we assume the following model:

$$Y_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{Z}_j\boldsymbol{\gamma}_j + \epsilon_j, j = 1, \ldots, m, \quad (3)$$

where Y_j is a numerical response, $\mathbf{X}_j$ and $\mathbf{Z}_j$ fixed matrices of covariates of rank $k > 0$ and $h \geqslant 0$ respectively (subject to $k+h < n$). $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ are vectors of parameters and ϵ_j is zero-centered and arising from a Gaussian random variable with second moment σ_j. The m error terms can be arbitrarily dependent. The null hypothesis to be tested is

$$H_{0j} : \boldsymbol{\beta}_j = \mathbf{0}. \quad (4)$$

allowing $\boldsymbol{\gamma}_j$ to be possibly different from zero. Let T_j be any test statistic associated with (4) that is ancillary to σ_j under the null. This setting includes Wald statistics, z, t and F tests, tests on correlations and on parameters of general linear models. Also many non-parametric tests can be seen to arise from a model of the kind (3), for instance through the normal score rank transformation (Conover and Iman, 1982).

We propose to order the hypotheses according to decreasing values of $M_j = Y_j'(\mathbf{I} - \mathbf{H}_j)'Y_j$, with $\mathbf{H}_j = \mathbf{Z}_j(\mathbf{Z}_j'\mathbf{Z}_j)^{-1}\mathbf{Z}_j'$. M_j is the second moment of residuals of the model, estimated constraining the parameters under the null hypothesis. Note that when $h = 0$, M_j simply reduces to n times the non-central second moment of Y_j, $Y_j'Y_j$ and we obtain a one-sample t or z test for (4). When Z_j is a vector of ones, M_j reduces to n times the central

second moment and we obtain two-sample t or z tests, and F tests. In cases in which there is a different sample size associated with different hypotheses, M_j can be divided by the corresponding sample size. We can now support our claim that hypotheses with a larger effect size are more likely put at the top of the list. This is precisely true under homoschedasticity. Suppose for simplicity that we are performing a univariate t-test. In that case, $M_j = Y_j' Y_j$ and therefore $E[M_j] \propto \sigma^2(1 + \mu_j^2)$. Hence, hypotheses with μ_j farther from zero will have a larger M_j and will be put at the top of the list. The same can be shown for the any other test statistic for verifying (4) under (3).

Next theorem shows that with our data driven ordering of hypotheses the FDR is still controlled at level q.

THEOREM 3: *Assume that model (3) holds, hypotheses (4) are under test. Denote test statistics T_j $\forall j = 1, \ldots, m$ and assume they are ancillary to σ_j under the null hypothesis. Compute p-values based on the upper tail of the null distribution of T_j. If the m hypotheses are tested along (decreasing) order of M_j, $j = 1, \ldots, m$, the FDR is controlled at level q under independence. If we use the modified threshold in Theorem 2, the FDR is controlled at level q under arbitrary dependence.*

A proof of the Theorem 3 is given in Web Appendix.

The rationale behind the proposed ordering is that large values of M_j will be associated with false null hypotheses, especially when the error variances are of the same order of magnitude. The reason is that under the null M_j will only be affected by σ_j, while under the alternative there will be an additional positive term due to location shift. We must here stress that when the error variance σ_j is constant or approximately constant over j the data driven ordering is particularly effective in prioritizing tests corresponding to false null hypotheses. On the other hand when variances are of a different order of magnitude, this effect is much less marked since variances will have an influence on the ordering.

4. Simulation study

We now study the performance of our algorithms through a brief simulation study. We perform one-sample t-tests, with data generated from standard normals under the null hypothesis. We let $n = 5, 10, 20, 50$; $m = 100, 1000, 10000$; $q = 0.01, 0.05, 0.10$. We fix the mean under the alternative hypotheses so that the single tests have a prescribed power of 70% when $q = 0.05$ and the proportion of false null hypotheses is fixed at 10%. Note that in this setting the single inference power is not bound to increase with the sample size, as the mean under the alternative decreases with n. We have chosen this setting since it clearly penalizes our approach with respect to competitors when the sample size is large. If we let the single inference power increase with the sample size, when n is large there are only minor differences among procedures. For each setting we generate the data, compute p-values, and apply the Benjamini and Hochberg (1995) (BH) and Benjamini and Yekutieli (2001) (BY) procedures; together with our procedure with data driven order of the hypotheses (Ord) and the version for general dependence (OrdGD). We perform 10000 Monte Carlo iterations and report the average fraction of correctly rejected hypotheses over the number of true alternative hypotheses available. We will refer to this fraction as "power" in this section, and to the marginal probability of correct rejection as "single inference power".

[Figure 1 about here.]

The fraction of correctly rejected hypotheses under independence is reported in Figure 1. The main conclusion from the simulations is that our procedure is particularly suited for the challenging cases in which the sample size is small. When $n = 5$, the Ord procedure is able to identify approximately five times as many false nulls than BH, and is better than any competitor up to $n < 20$ in many cases. This is particularly evident when $q = 0.01$, indicating that the procedure is also more suited for low single inference power cases. For what concerns procedures designed for general dependence, it can be seen that there is a slight loss of power of BY with respect ot BH (see also Farcomeni (2006)). The performance of OrdGD as compared to BY is along the same lines as before: OrdGD gives larger power than BY when n and/or q are small. rather surprising, outperforming the competitor BY in all cases and often achieving a power very close to that of the BH procedure, which is valid only under much more restrictive assumptions. An absolute effect of m is seen on power, but the ranking among procedures is approximately constant with m.

The fixed single inference power setting further underlines that, as n increases, the data driven order of hypotheses is less and less able to put false nulls at the beginning of the list. Roughly speaking, effect size is blurred by a larger sum of squares of the errors, when n is large. We have chosen to perform simulations under a fixed single inference power to underline this limitation of our approach, which is not seen if we let the single inference power increase with sample size.

Simulations so far are performed under independence assumptions. We now report on dependent test statistics, where data are generated from a multivariate normal in $\mathcal{R}^m$. In order to impose a meaningful dependence structure, we pretend data is scattered on a squared spatial grid. Random variables are then linked to an ideal map, equally scattered, and indexed by a pair of coordinates (i, j), i indicating the "row" position and j indicating the "column" position. In each point (i, j) a sample of n normals is observed. A similar setting is used in Farcomeni (2006, 2007). We determine the covariance matrix through a simplified version of a kernel commonly used in spatial statistics:

$$\mathrm{Cov}(X_{ij}, X_{i'j'}) = \cos\left(\frac{1}{\tau}\, d((i,j),(i',j'))\right) e^{-\frac{1}{\tau}\, d((i,j),(i',j'))}, \quad (5)$$

where $d(\cdot, \cdot)$ is the euclidean distance function, (i, j) are the coordinates of the h-th test and (i', j') the coordinates of the z-th test. This function is used to generate both positive and negative correlations. The strength of dependence

is controlled by the tuning parameter τ, which is set as $\tau = 7.5$, leading to correlations as high as 0.9. As noted by the AE, this *damped cosine dependence* scenario is close to a typical microarray setting.

Figure 2 and Figure 3 report the estimated power and the estimated level of each multiple test, respectively.

[Figure 2 about here.]

[Figure 3 about here.]

It can be seen that the nominal error rate is not exceeded even if we generated dependent data. There is generally a slightly smaller power, but the comparison among procedures is along the same lines of the independence case. It is worth noticing that when $q \geqslant 0.05$ and $n = 50$, surprisingly enough BY outperforms BH. We report other simulation studies under dependence in our online supplementary material.

We would like to underline that the findings of our simulation study suggest that prior ordering may be very useful in increasing power. We are aware it would be problematic in some cases, and that is the reason why we propose a data-driven ordering in addition to the a-priori ordering. What simulations suggest is actually not new in the FDR literature: Genovese et al. (2006) and Roeder and Wasserman (2009) use weighting rather than ordering, but obtain results that are along the same lines. They suggest in fact that prior weighting may significantly increase power, and in cases in which it does not the loss is negligible.

5. Application to genome wide aCGH analysis of intracranial ependymoma

We consider data collected by Modena et al. (2006) and freely available. In order to identify clinically relevant molecular signatures of intracranial ependymoma, a sample of patients with infratentorial (IT, $n = 14$) tumor location is compared with a sample of patients with supratentorial (ST, $n = 8$) tumor location. The 22 samples were analyzed with a 1-megabase resolution aCGH chip produced by spotting in triplicate 3,612 degenerate oligonucleotide primer polymerase chain reaction (PCR) amplified bacterial or P-1 derived artificial chromosomes (BAC/PAC) clones onto glass slides (more details are given in Modena et al. (2006)). We have clones in chromosomes 1 to 22, with a minimum of 56 in chromosome 21 and a maximum of 253 in chromosome 4.

The adequate test for the null hypothesis of no difference between IT and ST is a standard two sample t-test. The hypothesis that the errors are normally distributed was checked visually and using formal testing, and it is not rejected. With reference to model (3), the t-test arises from setting $\mathbf{Z}_j$ as a column of ones (i.e. the intercept) and $\mathbf{X}_j$ as an indicator of ST patients. The resulting H_j matrix is 22x22 matrix with elements all equal to $1/22$, and the resulting M_j is therefore the total sum of squares in each sample, that is, the sum of the squared differences between the sample expressions and their overall mean.

To give an idea of how our data driven procedure is working, we report Figure 4 which shows a scatterplot of the second central moment of each sample (which is proportional to M_j) vs $-\log_{10}(p_j)$, restricted to chromosome 9. The clear trend that can be seen in the figure indicates that the data driven ordering will likely enhance power.

[Figure 4 about here.]

[Table 2 about here.]

In Table 2 we compare the number of rejections for the BH, BY, Ord and OrdGD procedures, where in the last two we used a data driven order based on M_j as defined above. BH and BY procedures lead to no rejections for $q \leqslant 0.05$. When $q = 0.1$, BY still leads to no rejections while BH leads to selection of 10 spots. Our Ord procedures are clearly more powerful. We have evidence that most of our rejections are not false discoveries: most of the discoveries are in fact located in chromosome 9: 1 (100%) when $q = .01$, 22 (85%) when $q = .05$ and 34 (85%) when $q = .10$ for our Ord procedure, all of them with OrdGD. This finding is coherent with other results in the literature, which are based on independent data: Schneider et al. (2009) performed a microsatellite analysis on chromosome 9 and found that *"genetic aberrations were found significantly more often in supratentorial tumors than in tumors with infratentorial location"*. See also references therein. A similar result is also reported in Pfister et al. (2007). A comparison of the mean expression levels of selected clones (not reported) confirms this finding also in our data: all rejected hypotheses in chromosome 9 have a positive mean difference. Moreover Yao et al. (2011) report that *"deletions on chromosome 9 occur more frequently in supratentorial tumors"*. With our aCGH analysis paired with the Ord and OrdGD procedures we can say even more. In fact, the active probes selected are not only located in chromosome 9, but in particular in a very limited region of this chromosome: with $q = .05$ and the Ord procedure, we have 8 rejections in domain $9p21$, 1 in $9p22$ (also rejected with $q = 0.01$), 7 in $9p23$ and 4 in $9p24$. 25 out of 40 rejections are in domains $9p21$ to $9p24$ with $q = 0.1$ using the Ord procedure. All rejections obtained with OrdGD are in domains $9p21$ to $9p24$. Hence we can narrow down evidence of substantial differences between IT and ST not only to chromosome 9, but in particular to its short arm, in domains 21 to 24.

Supplementary material

Supplementary material is available at the Biometrics website on Wiley Online Library. It includes proofs of the formal results, results of simulation studies under different settings, as a complement of the simulation study of Section 4, and an additional real data example based on *a priori* order of the hypotheses.

Acknowledgments

The authors are grateful to an associate editor and three referees for detailed comments that helped clarify the proofs and the presentation. LF was supported by grant from the University of Padua (Progetti di Ricerca di Ateneo 2011,

project CPDA117517) and by the Cariparo Foundation Excellence grant 2011/2012.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society (Ser. B)* **57,** 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29,** 1165–1188.

Conover, W. and Iman, R. (1982). Analysis of covariance using the rank transformation. *Biometrics* **38,** 715–724.

Dmitrienko, A., Tamhane, A. C., and Wiens, B. L. (2008). General multistage gatekeeping procedures. *Biometrical Journal* **50,** 667–677.

Dudoit, S., Shaffer, P., and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18,** 71–103.

Farcomeni, A. (2006). More powerful control of the false discovery rate under dependence. *Statistical Methods & Applications* **15,** 43–73.

Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scandinavian Journal of Statistics* **34,** 275–297.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research* **17,** 347–388.

Finos, L. (2011). A note on left-spherically distributed test with covariates. *Statistics & Probability Letters* **81,** 639 – 641.

Finos, L. and Farcomeni, A. (2011). k-FWER control without p-value adjustment, with application to detection of genetic determinants of multiple sclerosis in Italian twins. *Biometrics* **67,** 174–181.

Genovese, C., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika* **93,** 509–524.

Hommel, G. and Kropf, S. (2005). Tests for differentiation in gene expression using a data-driven order or weights for hypotheses. *Biometrical Journal* **47,** 554–562.

Hsu, J. and Berger, R. (1999). Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *Journal of the American Statistical Association* **94,** 468–475.

Kirk, R. (2007). Effect magnitude: a different focus. *Journal of Statistical Planning and Inference* **137,** 1634–1646.

Kropf, S. and Läuter, J. (2002). Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. *Biometrical Journal* **44,** 789–800.

Kropf, S., Läuter, J., Eszlinger, M., Krohn, K., and Paschke, R. (2004). Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses. *Journal of Statistical Planning and Inference* **125,** 31–47.

Läuter, J., Glimm, E., and Kropf, S. (1998). Multivariate tests based on left-spherically distributed linear scores. *Annals of Statistics* **26,** 1972–1988.

Marcus, R., Peritz, E., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63,** 655–660.

Maurer, W., Hothorn, L., and Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In Vollman, J., editor, *Biometrie in der chemischepharmazeutichen Industrie*, volume 6, Stuttgart. Fischer Verlag.

Modena, P., Lualdi, E., Facchinetti, F., Veltman, J., Reid, J. F., Minardi, S., Janssen, I., Giangaspero, F., Forni, M., Finocchiaro, G., Genitori, L., Giordano, F., Riccardi, R., Schoenmakers, E. F., Massimino, M., and Sozzi, G. (2006). Identification of tumor-specific molecular signatures in intracranial ependymoma and association with clinical characteristics. *Journal of Clinical Oncology* **24,** 5223–5233.

Pfister, S., Remke, M., Toedt, G., Werft, W., Benner, A., Mendrzyk, F., Wittmann, A., Devens, F., von, Hoff, K., Rutkowski, S., Kulozik, A., Radlwimmer, B., Scheurlen, W., Lichter, P., and Korshunov, A. (2007). Supratentorial primitive neuroectodermal tumors of the central nervous system frequently harbor deletions of the CDKN2A locus and other genomic aberrations distinct from medulloblastomas. *Genes Chromosomes Cancer.* **46,** 839–851.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical Science* **24,** 398–413.

Rosenbaum, P. (2008). Testing hypotheses in order. *Biometrika* **95,** 248–252.

Schneider, D., Monoranu, C., Huang, B., Rutkowski, S. Gerber, N., Krauss, J., Puppe, B., and Roggendorf, W. (2009). Supratentorial ependymomas show more frequent deletions on chromosome 9 than infratentorial ependymomas: a microsatellite analysis. *Cancer Genet Cytogenet* **191,** 90–96.

Strassburger, K., Bretz, F., and Finner, H. (2007). Ordered multiple comparisons with the best and their applications to dose-response studies. *Biometrics* **63,** 1143–1151.

Yao, Y., Mack, S. C., and Taylor, M. D. (2011). Molecular genetics of ependymoma. *Chinese Journal of Cancer* **30,** 669 – 680.

Received . *Revised* .
Accepted .

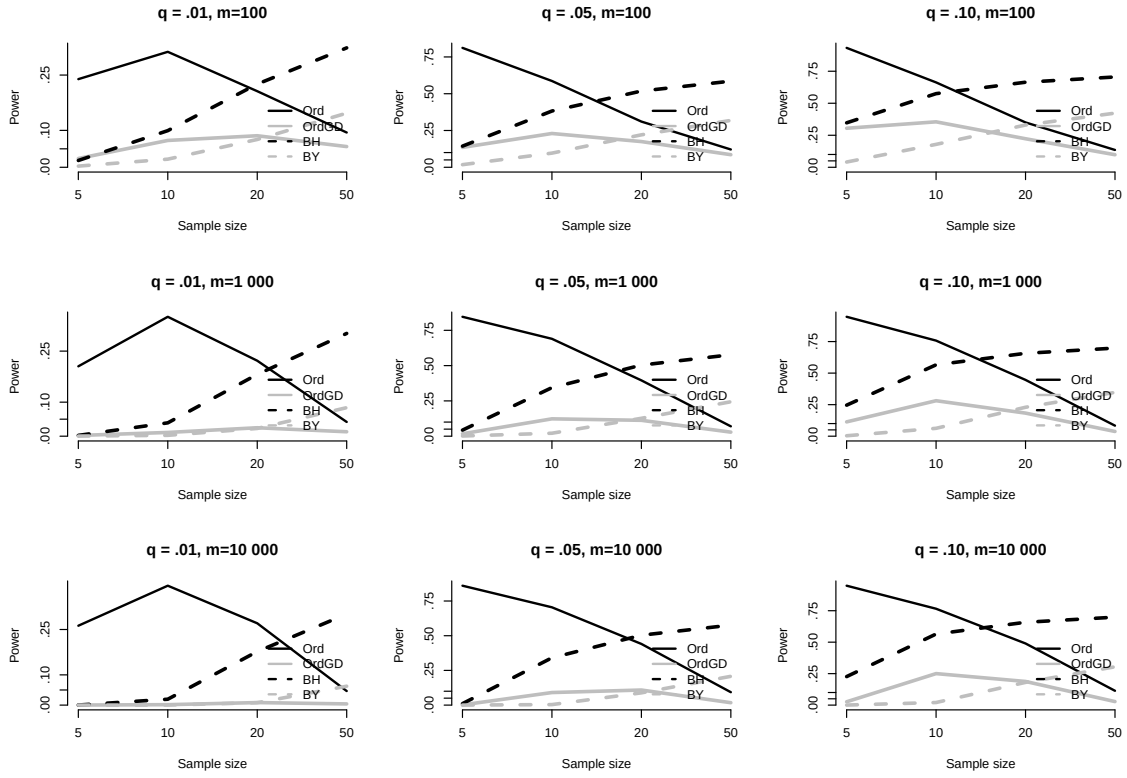

Figure 1. Proportion of correctly rejected hypotheses for different values of q, n, m. The proportion of false nulls is set at 10%, the power of each single test at 70%. Results are based on $B = 10000$ iterations.

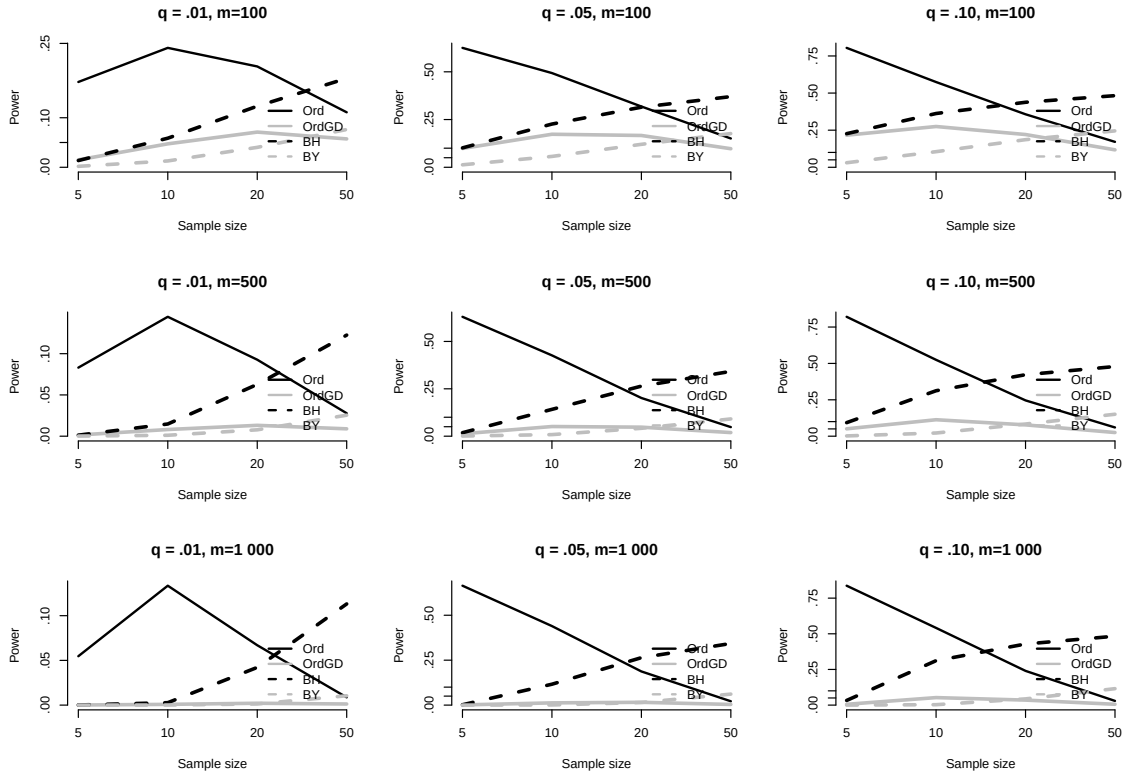

Figure 2. Proportion of correctly rejected hypotheses for different values of q, n, m. The proportion of false nulls is set at 10%, the power of each single test at 70%. Damped cosine dependence.

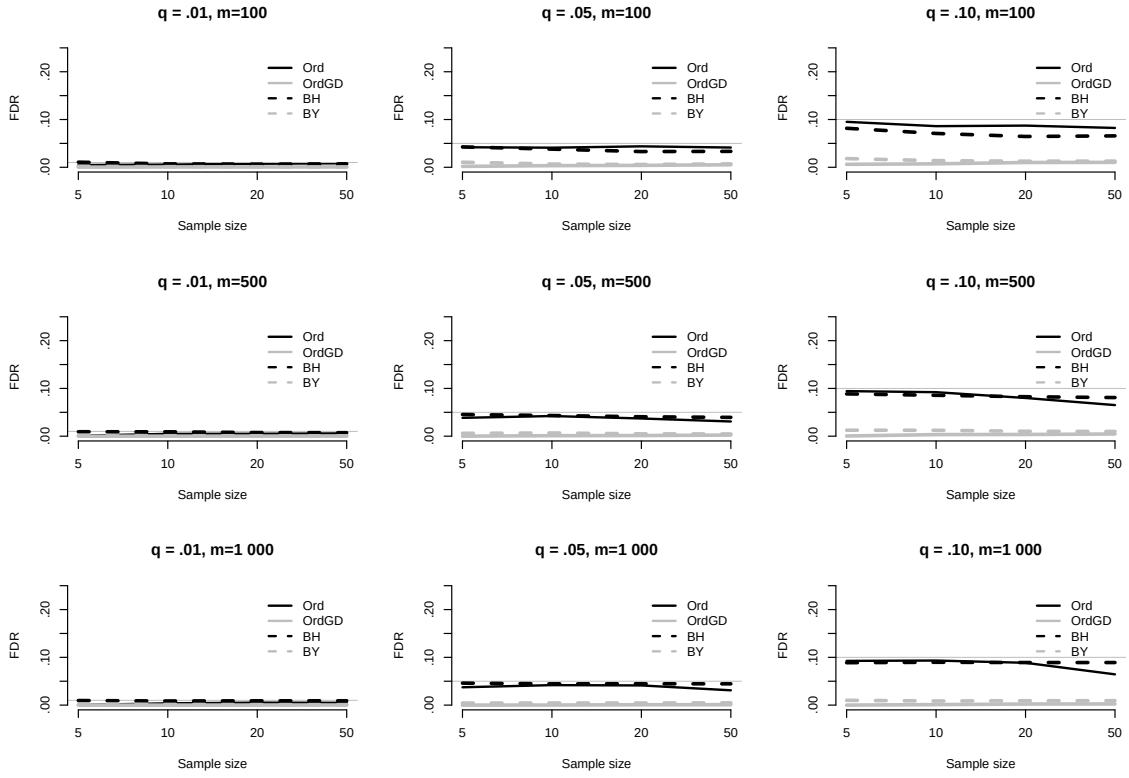

Figure 3. FDR for different values of q, n, m. The proportion of false nulls is set at 10%, the power of each single test at 70%. Damped cosine dependence.

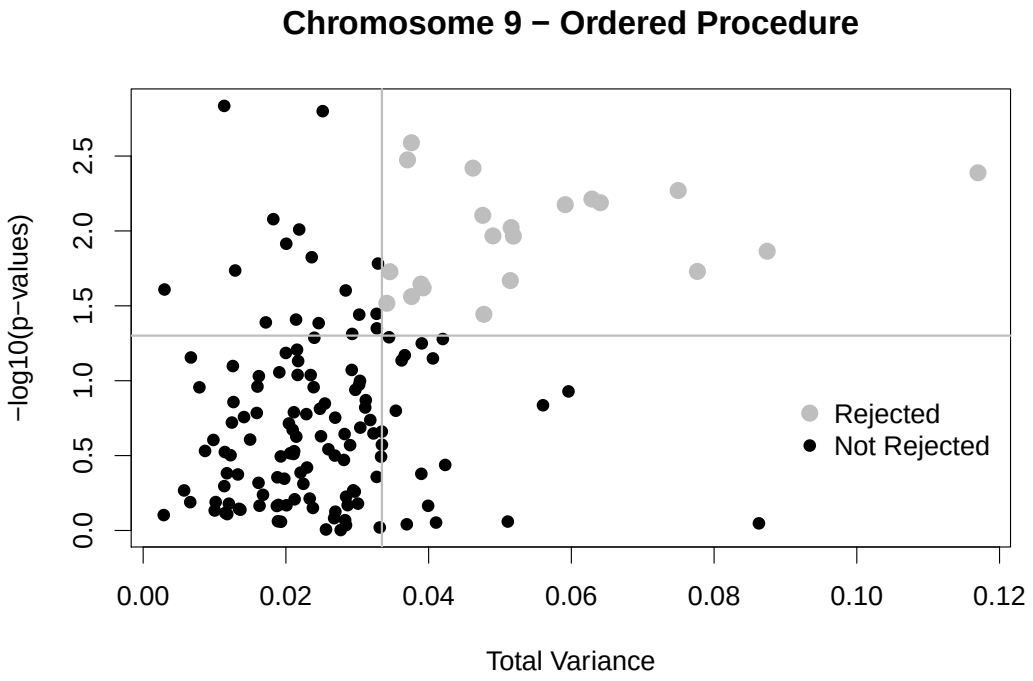

Figure 4. Scatter plot of second moment vs $-\log_{10}$ p-value for intracranial ependymoma data. The second moment is used later for ordering p-values.

Table 1
Outcome in testing m hypotheses

	Null Hypotheses		
	Not Rejected	**Rejected**	**Total**
True	$N_{0\|0}$	$N_{1\|0}$	M_0
False	$N_{0\|1}$	$N_{1\|1}$	M_1
Total	$m - R$	R	m

Table 2

Number of rejections for the intracranial ependymoma data with different multiple testing procedures and significance levels q.

	BH	BY	Ord	OrdGD
$q = .01$	0	0	1	0
$q = .05$	0	0	26	5
$q = .10$	10	0	40	16